Proposal for Upper Division Course in Neural Data Science

Justification for course

Neuroscience is a rapidly changing field that is increasingly moving towards bigger datasets and computational approaches. The field of data science is concurrently evolving across numerous disciplines. Regardless of the specific discipline, success in working with such large datasets relies on the ability to translate data into meaningful observations. With the creation of the Halıcıoğlu Data Science Institute at UC San Diego, there are increasing opportunities to bridge data science with other fields. In particular, there is a strong need to neuroscientists who can think deeply about problems in data science, including parameterizing variables, working with large datasets, and handling multiple data types. Fields such as genomics have successfully developed curricula and resources that allow students to access, improve, and conduct research with big data—we plan to develop such curricula for neuroscience. This interdisciplinary, project-based course will provide upper division Cognitive Science and Data Science students at UC San Diego with practical experience for future careers in research, neuroscience, and other fields of data science.

Course title and description

*Neural Data Science (4U)*

Project-based course in which students will use computational notebooks to perform exploratory data analyses and to test hypotheses in large neuroscience datasets, including the differences between unique neuron types, leveraging text mining of the neuroscience literature, and human neuroimaging analyses.

Course logistics

This upper division 4-unit course is designed for all Cognitive Science or Data Science students with upper division standing. Bradley Voytek will teach this course through Cognitive Science while Ashley Juavinett will teach it through Biological Sciences. Voytek and Juavinett will keep separate course numbers within their respective departments (Cognitive Science and Biology), but will collaborate to develop this course together, teaching it in separate quarters.

Given the increasing interest in large scale analytics, and students joining the new Data Science program, we expect that this course will attract a substantial number of students each quarter. It will be designed to accommodate 200 students, and would ideally have access to computer lab space. The course will require the same resources as other upper division cognitive science and data science electives (classroom with standard projection or screen technology, equipped for podcasting, and Instructional Assistants).

Neural Data Science will be a lecture and project-based course. It will consist of three hours of lecture (two 1.5 hour sessions per week); one that is designated for lecture, and one designated for in-class project work and discussion. Homework will consist of problem sets designed to support students with their progression through the larger class projects. Students will be evaluated on their in-class work and participation (20%), problem sets (20%), and three large projects (20% each). The final exam will be in the form of a final project.

* Proposed course code: COGS 138

Pre-requisites

* COGS 18: Introduction to Python
* COGS 118C: Neural Signal Processing

Tentative syllabus

**Students will be able to**:

* *Develop* hypotheses specific to big data environments
* *Choose* a model based on the underlying data
* *Transform* data into meaningful parameters for statistical analysis
* *Design* a big data experiment and *excavate* data from open sources
* *Consider* alternative hypotheses for trends in the data

**Topics covered**

*Weeks 1 & 2: Tools for data science*

To set the foundation for this course, we’ll introduce tools that are commonly used to analyze big data sets in neuroscience, including Jupyter Notebooks and SDKs, as well as Python and relevant packages (MNE, NLTK, etc.). The first two weeks will serve to bring students up to speed on the basics of using these tools to analyze neural data.

Weeks 3 & 4: *Time series analysis and signal processing*

How do we perform analyses on open source datasets for large-scale electrophysiology (single cell, EEG, and MEG) and imaging (two-photon) data? Students will perform large-scale time series analyses and learn about the various types of signal processing used for different data types, and cloud-based, parallel analysis. This will extend topics covered in COGS 118C.

Weeks 5 & 6: *Parameterizing heterogeneous datasets*

Modern neuroscience incorporates various types of data, both physiological and behavioral. This portion of this course will address how we integrate diverse types of data (e.g., analog signals, video, text data, etc.) to address an experimental question. May work on recapitulating existing large-scale neuroscience text analyses such as NeuroSynth and brainSCANr.

Weeks 7 & 8: *Challenges and opportunities*

With larger data sets come newer challenges and opportunities. It is becoming more common to avoid signal averaging analyze signa-trial data. Here, we’ll explore different approaches to investigating single trials of population data, such as Markov models and various regression analyses, as well as the challenges in algorithmic data pre-processing, cleaning, feature selection, and dimensionality reduction.

Weeks 9 & 10: *Final projects*

In the last two weeks of the course, students will work on final projects in which they use any of the open source datasets that we have encountered to address pressing questions in neuroscience, such as the network properties of the connectome, the role of different cell types in the cortex, or the cortical representation of language.